# Bokomtaler

## Gisle Andersen (red.): *Exploring Newspaper Language. Using the web to create and investigate a large corpus of modern Norwegian.* Amsterdam: John Benjamins 2012, 362 s.

Utgangspunktet er prosjektet Norsk Aviskorpus (Norwegian Newspaper Corpus, NNC). Som navnet tilsier, består korpuset av norske avistekster som er samlet inn fra 1998 til i dag. Boken inneholder artikler som beskriver korpuset, og eksempler på hvordan det kan brukes i språkvitenskapelig forskning.

Store tekstmengder åpner for nye typer språkstudier. I språkvitenskapelig sammenheng kan man ved hjelp av slike samlinger identifisere nyord og oppdage nye brukskontekster for ord eller uttrykk, leksikografer kan ekstrahere ord som bør inn i ordbøker, eller legge nye bruksmåter til eksisterende ord, språkteknologer kan derivere statistiske språkmodeller til bruk i f.eks. talesyntese eller maskinoversettelse, osv. Som påpekt av Gisle Andersen og Knut Hofland i forordet er avistekster spesielt godt egnet til å bygge en omfattende norsk tekstsamling fordi de har en rimelig balansert tematisk spredning, de er skrevet av profesjonelle skribenter, og de dekker en etter hvert nokså lang periode. Og sist men ikke minst, avistekstene publiseres jevnlig på nett, noe som gjør at dataprogrammer kan laste dem ned daglig uten at utgiveren trenger å bruke ressurser på innsamlingen.

NNC inneholder mer enn en milliard løpende ord, noe som gjør det til et stort korpus som tåler internasjonal sammenligning. British National Corpus regnes som et stort korpus og inneholder ca. en tidel av NNC (BNC er riktignok mer balansert med hensyn til genre enn NNC). Tekstene i NNC er kategorisert etter tema, målform, utgiver, tid og skribent. Hver tekst er dessuten automatisk tagget (merket) med hensyn til grammatiske egenskaper som ordklasser, kjønn, form, tid osv. Dermed er det mulig å søke etter grammatiske mønstre sammen med ordformer, om man er interessert i det.

Det er en langt fra enkel oppgave å luke vekk uønskede elementer fra tekstene som skal inkluderes i korpuset, f.eks. web-adresser, html-koder, reklamesnutter og lignende. Prosjektet har over lang tid opparbeidet erfaring som gjør at undertegnede føler seg trygg på at irrelevante egenskaper ved tekstene er fjernet. Et beslektet, men mer komplisert problem, er parallell publisering av identiske tekster, f.eks. at flere utgivere kopierer artikler, eller deler av artikler, fra andre aviser. Dermed risikerer man at samme tekst opptrer flere ganger i korpuset. Prosjektet har inkludert slike republiseringer såfremt de opptrer i ulike publikasjoner. Dette er uheldig fordi man risikerer å få et feilaktig inntrykk av konteksten til lavfrekvente ord eller fraser som opptrer i slike republiserte tekster. Korpuset er uansett så stort at det å fjerne dem ikke vil forringe det i nevneverdig grad.

Boken er organisert i to deler, der den første dreier seg om metoder og verktøy, mens den andre fokuserer på korpusbaserte case-studier. Artiklene blir behandlet kronologisk.

På grunn av rettighetsavtalene med utgiverne kan man bare få indirekte tilgang til NNC via søkeverktøy som viser setninger eller fraser. Kvaliteten til søkeverktøyene er derfor avgjørende for hvilken verdi korpuset har for dem som vil bruke det. Når et korpus blir så stort som NNC, risikerer man at brukeren må vente uforholdsmessig lenge før søkeprogrammet returnerer resultatet av et søk. Problemet blir forsterket av at NCC også er annotert med grammatisk informasjon som også skal være søkbar. For å løse slike problemer har Paul Meurer utviklet en imponerende plattform (Corpuscle) som kan brukes til å organisere og tilgjengeliggjøre annoterte korpuser for avanserte søk, deriblant NNC. Systemet er beskrevet i artikkelen *Corpuscle - a new corpus management platform for annotated corpora*. Det tillater utstrakt bruk av regulære uttrykk, noe som gjør at man kan formulere detaljerte og målrettede søk og få resultatene servert usedvanlig raskt tilbake. Brukere som ikke er fortrolige med regulære uttrykk, tilbys et menybasert grensesnitt, noe som er velkomment for språkforskere med liten kjennskap til databehandling og slike regulære uttrykk. Uten systemer som Corpuscle blir det vanskelig å studere informasjonen som finnes i store annoterte korpuser. Systemet er også i bruk for andre korpuser, bl.a. Talebanken i tilknytning til Norsk Dialektkorpus og INESS (se nedenfor).

I artikkelen *OBT-stat* viser Janne Bondi Johannessen, Kristin Hagen, André Lynum og Anders Nøklestad hvordan man kan entydiggjøre en regelbasert ordklassetagger slik at den bare returnerer *en* grammatisk beskrivelse for hvert ord som analyseres. Automatisk ordklassetagging er en forenklet form for grammatisk analyse der et dataprogram tilordner hvert ord grammatisk relevant informasjon som ordklasse, bestemthet, form, genus, numerus osv. Ordklassetaggere er enten statistisk basert, det vil si at de er trent opp på en samling av

manuelt kontrollerte setningsannotasjoner, eller de er regelbaserte, som Oslo-Bergen-taggeren. Andre varianter finnes også, uten at vi skal gå inn på dem her. For norsk er Oslo-Bergen-taggeren mest brukt av språkforskere, rett og slett fordi den gir mest interessant grammatisk informasjon. Det kan innvendes mot regelbaserte taggere at de ikke alltid klarer å velge én og bare én korrekt analyse av alle ord i setningene den analyserer, selv om det av og til ikke er mulig å bestemme hvilken analyse som er korrekt. Et eksempel er setninger som *Neste prosess er å skrive bok* dersom ordet *bok* er oppført som *substantiv, ubestemt, entall, hankjønn* og som *substantiv, ubestemt, entall, hunkjønn*. Slike flertydighetsproblemer kan unngås dersom *bok* beskrives entydig som *substantiv, ubestemt, entall, hankjønn/hunkjønn*, men flertydigheten er fremdeles til stede. Johannessen *et al.* demonstrerer hvordan slike flertydigheter kan løses opp ved å la en standard statistisk basert tagger bestemme hvilken tagg som skal velges dersom den regelbaserte taggeren etterlater seg flertydigheter. Den statistiske taggeren er trent opp på materiale som er manuelt annotert med samme type grammatisk informasjon som Oslo-Bergen-taggeren benytter. Evalueringen viser at OBT-stat oppnår svært gode resultater sammenlignet med andre taggere (ca. 96 % korrekte tagger). Spesielt overbevisende er det at OBT-stat benytter et adskillig rikere utvalg tagger enn det som er vanlig i automatisk ordklassetagging. Med OBT-stat kan hele aviskorpuset ordklassemerkes, og språkforskere kan dermed søke etter grammatiske mønstre i en gedigen tekstsamling.

Victoria Rosén diskuterer i artikkelen *Exploring corpora through syntactic annotation* hvilke muligheter et rikt syntaktisk annotert korpus gir språkforskere som ønsker å studere syntaktiske fenomener i detalj. En ordklassetagger, jf. artikkelen til Johannessen *et al.*, bidrar ikke med strukturell grammatisk informasjon. Dermed kan man ikke søke etter f.eks. relativsetninger. Rosén viser hvordan man på delvis automatisk vis kan annotere en ressurs som NNC med rikholdig syntaktisk og semantisk informasjon. Dette er hovedoppgaven til prosjektet INESS (Infrastructure for the Exploration of Syntax and Semantics). Setningene blir analysert automatisk ved hjelp av en parser og en maskinlesbar grammatikk, i dette tilfellet grammatikken til Norgram-prosjektet ved Universitetet i Bergen. Dersom grammatikken returnerer flere analyser til en setning, må man velge den riktige analysen. Skulle den riktige analysen ikke være blant de som presenteres, eller dersom grammatikken ikke klarer å analysere en setning, må grammatikken modifiseres. Rosén argumenterer, overbevisende, for at en slik ressurs med entydig syntaktisk analyserte setninger vil være verdifull for språkforskere og datalingvistisk forskning og utvikling.

Gunn Inger Lyse og Gisle Andersen diskuterer i artikkelen *Collocations and statistical analysis of n-grams* hvordan man kan identifisere flerordsuttrykk i et stort korpus som NNC. Flerordsuttrykk forstås her som sekvenser av to eller tre ord, altså bigrammer og trigrammer, som har leksikalsk, syntaktisk,

semantisk eller pragmatisk *fasthet* (idiomaticity). Eksempler er *de facto* og *i motsetning til*, mens *noen mener* eller *det kan kanskje* ikke er flerordsuttrykk, selv om de er eksempler på bigrammer og trigrammer. Forfatterne påpeker at flerordsuttrykk er viktige å registrere i leksikografisk og terminologisk arbeid, og at det er svært viktig å være oppmerksom på dem i ordklassetagging og automatisk setningsanalyse, nettopp fordi de har egenskaper man ikke forventer når ordene betraktes isolert. Men det er vanskelig å finne målrettede metoder som plukker ut alle flerordsuttrykk fra et stort tekstkorpus på noenlunde presist vis, og dette er hovedtemaet i artikkelen. Forfatterne har brukt til sammen ni ulike strategier og evaluerer deres evne til å identifisere anglisismer, andre uttrykk fra fremmede språk (f.eks. *gefundenes fressen*), fagtermer, idiomatiske uttrykk, etc. Resultatene er ikke enkle å fortolke, men artikkelen viser at noen metoder er å foretrekke for visse typer flerordsuttrykk (f.eks. anglisismer), mens andre metoder passer til andre typer (f.eks. fagtermer). Uansett er det påkrevet at man går gjennom resultatene manuelt og siler ut irrelevante forslag.

I artikkelen *Automatic topic classification of a large newspaper corpus* viser Thomas Hagen hvordan man automatisk kan foreta emneklassifikasjon av tekster i et stort korpus, dvs. om tekstene handler om kultur, underholdning, utenriks, politikk, sport, etc. Fremgangsmåten er at et maskinlæringsprogram ekstraherer egenskaper som er typiske for de ulike emnene fra et treningskorpus som er manuelt klassifisert. Et grunnleggende problem i emneklassifisering er at emnene er overlappende, og det viser seg at menneskelige annotører ikke klarer å gjøre det riktig i mer enn 59 % av tilfellene. Hagen har eksperimentert seg frem til en metode som, forenklet sagt, går ut på å først identifisere artikler som handler om sport, deretter de som handler om innenriksstoff, så økonomi, vitenskap og teknologi, politikk osv. Systemet hans klarer å klassifisere korrekt i 54 % av tilfellene, noe som må regnes som godt siden humanannotører ikke klarer bedre enn 59 %.

Gyri Smørdal Losnegaard og Gunn Inger Lyse tar i artikkelen *A data-driven approach to anglicism identification in Norwegian* for seg automatisk identifikasjon av anglisismer i norske tekster, som *podcasting*, *project manager* osv. Leksikografer ønsker å identifisere anglisismer i norsk på et tidlig stadium, både for å få dem registrert og eventuelt se etter mulige norske avløserord. Identifikasjonen skjer, forenklet formulert, ved at et dataprogram lærer seg entydige engelske og norske sekvenser av tre tegn, som *ect*, *oph* osv, og disse brukes til å finne ord som programmet regner som engelske eller norske. Det gjøres tester mot ulike typer data, men motivasjonen for de ulike datasettene er uklar for undertegnede. De mest innsiktsfulle testene må være mot en uavhengig "gullstandard" bestående av håndplukkede engelske og norske ord. Det viser seg at metodene er i stand til å identifisere ca. en tredel av anglisismene, og ca. 40 % av de postulerte anglisismene er korrekte. Det er vanskelig å vurdere om dette

er et godt resultat, og forfatterne trekker ingen bastante konklusjoner utover å påpeke at metodene vil være et nyttig supplement til andre metoder.

Artikkelen til Losnegaard og Lyse avslutter første del av boken. Oppsummerende vil jeg si at artiklene i denne delen har noe varierende kvalitet og interesse. Meurers artikkel er mest innovativ og presenterer et nytt og svært interessant verktøy for arbeid med annoterte korpuser. Med unntak av artikkelen til Johannessen *et al.* har de øvrige artiklene et preg av at man ønsker å si noe om hvordan man kan eksperimentere med og bruke NNC, uten at det egentlig presenteres nye innsikter. Del to har et klarere empirisk tilsnitt, og her finner man de mest interessante artiklene.

Gisle Andersens *A corpus-based study of the adaption of English import words in Norwegian* kan gjerne betraktes som et mønsterarbeid for hvordan man kan bruke et stort aviskorpus til å teste hypoteser og ekstrahere empiriske generaliseringer. Utgangspunktet er Språkrådets forslag til avløserord for anglisismer fra 1996 og 2004. Andersen studerer både i hvilken grad de foreslåtte avløserordene har slått rot, og hva som kan være årsakene til at kun noen ganske få av dem har festet seg. Ved hjelp av Corpuscle (se ovenfor) ekstraheres alle de foreslåtte avløserordene fra NNC, om de finnes der. Ord som *keitering*, *overhedd* og *sjarter* forekommer ikke, ord som *innputt* og *utputt* forekommer, men de engelske ordene *input* og *output* er nesten enerådende. Derimot er ord som *ketsjup* og *pønk* mer vanlige, mens *rapp*, *klinsj* og *streit* brukes langt mer enn de tilsvarende engelske ordene. I siste del av artikkelen presenterer Andersen noen generaliseringer som ser ut til å påvirke i hvilken grad norske avløserord slår an, f.eks. at stor ortografisk forskjell er ugunstig for norske avløserord, mens mindre forskjeller har motsatt effekt. Generaliseringenes holdbarhet kan diskuteres, men Andersen argumenterer overbevisende for at tendensene er til stede.

*Norm clusters in written Norwegian* er skrevet av Helge Dyvik, og denne artikkelen viser hvordan man ved hjelp av korpusbasert metodikk kan dokumentere implikasjonsmønstre i bokmål og nynorsk. Med *implikasjonsmønster* forstår man at dersom en skribent bruker *-en* i bestemt form entall av feminine substantiver som *hytte*, så vil vedkommende også bruke endelsen *-et* i bøyningen av verb som *kaste*. Men implikasjonen gjelder ikke andre veien, dvs. at en som foretrekker *kastet*, ikke nødvendigvis også skriver *hytten*. Derimot vil en som skriver *kasta*, foretrekke *hytta*, men dette er heller ikke en toveis implikasjon. Man ender dermed opp med to bøyningsmønstre som gjensidig utelukker hverandre, altså *kasta* og *hytten*, mens *kastet* og *hytta* fremstår som typiske og nøytrale. Spørsmålet som Dyvik tar tak i, er om dette bare er noe man kan ha en intuitiv forståelse av, eller om det kan belegges empirisk. Han benytter seg av ordlister fra Norsk Ordbank og ordklassemerkede tekster fra NCC (jf. Johannessen *et al.* ovenfor) som er kategorisert etter publikasjon og forfatter.

Ved hjelp av korrespondanseanalyse av valgfrie ordformer kan man i et todimensjonalt rom identifisere perifere former og mer sentrale former. Implikasjonene går fra de perifere formene til de sentrale, men ikke omvendt. Analysen munner ut i at noen kjente tendenser blir verifisert, f.eks. at der det er valg mellom feminine og maskuline substantiver, foretrekkes normalt den maskuline, og at de som bruker a-form av svake verb, også bruker feminin bøyning av abstrakte begreper (*forestillinga* etc.). Metoden som benyttes i denne artikkelen, bør kunne brukes til å verifisere eller identifisere andre egenskaper som kjennetegner subnormer i de to skriftspråkene våre. Dermed kan den bli et meget verdifullt bidrag til empirisk basert normering av norsk skriftspråk.

Ruth Vatvedt Fjeld og Lars Nygaard gir i artikkelen *Lexical neography in modern Norwegian* en kort oversikt over leksikalsk neografi i Norge, og de viser hvordan man kan identifisere neologismer ved hjelp av korpuser som NNC. De gir en kort og konsis oversikt over leksikografisk behandling av neologismer i norsk, og argumenterer for at neologismestudier må ta for seg nyord ("neoformatives"), nye betydninger av eksisterende ord ("neosemanticisms") og nye fraser knyttet til eksisterende ord ("neophrasemes"). Identifisering av nyord kan være svært tidkrevende, også dersom man har tilgang til lister av nye ord som registreres i tekstsamlinger som NNC. Man ønsker å unngå "leksikalske døgnfluer", og Fjelds og Nygaards strategi for å identifisere nyord over tid er et nytt bidrag til leksikografisk metodikk. På denne måten kan man både identifisere stabile nyord som tas i bruk, og samtidig registrere slangord eller sammensetninger som etablerer seg i skriftspråket over tid. Slik longitudinell verifikasjon er et svært nyttig redskap, som bør brukes sammen med eller i tillegg til tradisjonell leksikografisk ekserpering. Dermed kan man unngå at registreringen blir tilfeldig, og kvaliteten til ordbøker blir forbedret.

Koenraad De Smedt behandler i artikkelen *Ash compound frenzy* den utstrakte bruken av sammensatte ord med forleddet *aske-* etter vulkanutbruddet på Island våren 2010, jf. eksempler som *askefast*, *askesky*, *askekaos* osv. Studier som dette er mulig fordi artiklene i NNC er datomerket. Dermed kan man søke etter ord som begynner med *aske* i et eller flere avgrensede tidsrom. Undersøkelser som dette gir nyttige bidrag til hvordan man kan studere produktive sammensetninger i detalj når man har tilgang til korpuser som er organisert som NNC.

I artikkelen *Financial jargon in general newspaper corpus* undersøker Marita Kristiansen bruken av visse finansbegreper (f.eks. *subprime* og *hedge fund*) i norske avistekster etter finanskrisen fra ca. 2007 og senere. Det viser seg at selv om den norske termen *råtne boliglån* også brukes, er *subprime*, kanskje noe overraskende, den varianten som foretrekkes. Kristiansen mener dette kan skyldes at *subprime* fungerer bra i sammensetninger (*subprimelån*,

*subprimeboble*, osv.), noe som virker plausibelt. Termen *hedge fund* omtales i de fleste tilfeller som *hedgefond*, altså med engelsk forledd og norsk etterledd, og man finner verbvarianter som *å hedge* eller *å hedge seg*. Bruken av engelske termer som *hedge* i norsk bør studeres nærmere, både med hensyn til sammensetninger og i andre sammenhenger. Kristiansen konkluderer med at korpusstudier av denne typen gir oss fersk terminologisk relevant informasjon som man ikke finner i lærebøker.

Metonymisk utvidelse og vaghet er tema for Sandra Halversons artikkel *Metonymic extension and vagueness*, med et spesielt blikk på *Schengen* og *Kyoto*. Disse bynavnene brukes også til å referere til avtalene som ble inngått der, medlemskap og geografisk område. Langackers *Cognitive Grammar* brukes som teoretisk grunnlag. Halverson viser empirisk hvordan stedsnavnene gradvis er blitt semantisk beriket. Korpuser som NNC er åpenbart en svært verdifull ressurs for denne typen studier.

Leiv Egil Breivik og Toril Swan analyserer i artikkelen *Spatial metaphors in present-day Norwegian* hvordan høy/lav-dimensjonen fra kognitiv lingvistisk teori manifesteres i norsk. Det viser seg, overraskende for undertegnede, at metaforisk bruk av *høy* og *høyere* utgjør mer enn 92 % av tilfellene som er undersøkt fra NNC. Forfatterne finner også en sterk tendens til metaforisk bruk av verbene *løfte* og *stige*, der 88 % av eksemplene fra NNC dokumenterer metaforisk bruk, og de finner også at denne opp-dimensjonen for det meste brukes i positiv kontekst. For lav-dimensjonen finner man omtrent samme andel metaforisk bruk for *lav*, *lavere*, *senke* og *synke*, og disse ordene brukes for det meste negativt, som teorien forutsier. Den høye andelen metaforisk bruk i avistekster er overraskende, og derfor er dette arbeidet også et strålende eksempel på hvordan NNC kan brukes til å ekstrahere ny kunnskap om norsk.

Den siste artikkelen i boken er Øivin Andersens *Doing historical linguistics using contemporary data*, der forfatteren ønsker å undersøke utviklingen til substantiver som er avledet fra verb (*skudd*, *slag*, osv.). De empiriske hypotesene som Andersen formulerer, basert på tidligere arbeider, lar seg dessverre ikke teste i et historisk perspektiv, rett og slett grunnet lite data fra perioden før 1990.

Artiklene i denne boken er av noe varierende kvalitet og interesse. Bokens innledning og artiklene til Meurer, Johannessen *et al.*, Gisle Andersen, Dyvik samt Breivik og Swan holder etter undertegnedes syn et høyt faglig nivå og gjør at boken bør være av stor interesse for alle som er opptatt av korpusrelatert språkvitenskap i Norge.

Torbjørn Nordgård
Universitetet i Nordland/
Lingit AS, Trondheim
torbjorn.nordgard@lingit.no